

W0082-01 EU

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-244526

(43)Date of publication of application : 08.09.2000

(51)Int.Cl.

H04L 12/28

G06F 13/00

H04L 29/14

(21)Application number : 11-044131

(71)Applicant : HITACHI LTD

(22)Date of filing : 23.02.1999

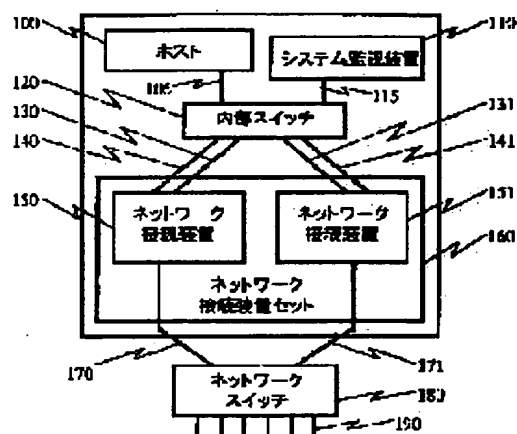
(72)Inventor : SAKURABA TATSUTOSHI  
SERIZAWA HAJIME  
TOMITA KAZUYOSHI

## (54) MULTIPLEXED NETWORK CONNECTOR SYSTEM

## (57)Abstract:

**PROBLEM TO BE SOLVED:** To utilize a duplexed network connector for improving a performance by distributing a load at the time of normal operation and to maintain usability without interposing any software of a host computer in the case of fault.

**SOLUTION:** At the time of reception processing, respective network connectors 150 and 151 are connected to a network switch 180 to receive the same reception data. The reception data to be transferred to a host are selected based on contents of the data, so that the reception data can not be missed and overlapped as a whole. In this case, when any fault occurs, the reception data to be selected by the remaining network connector are dynamically changed so as not to lose a reception function as a whole.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]



(19) 日本国特許庁 (J P)

## (12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2000-244526

(P2000-244526A)

(43) 公開日 平成12年9月8日(2000.9.8)

| (51) Int.Cl. <sup>7</sup> | 識別記号  | F I           | テーマコード <sup>*</sup> (参考) |
|---------------------------|-------|---------------|--------------------------|
| H 0 4 L 12/28             |       | H 0 4 L 11/00 | 3 1 0 D 5 B 0 8 9        |
| G 0 6 F 13/00             | 3 5 1 | G 0 6 F 13/00 | 3 5 1 M 5 K 0 3 3        |
| H 0 4 L 29/14             |       | H 0 4 L 13/00 | 3 1 1 5 K 0 3 5          |

審査請求 未請求 請求項の数 1 O L (全 13 頁)

(21) 出願番号 特願平11-44131

(22) 出願日 平成11年2月23日(1999.2.23)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 櫻庭 健年

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72) 発明者 芹沢 一

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(74) 代理人 100068504

弁理士 小川 勝男

最終頁に続く

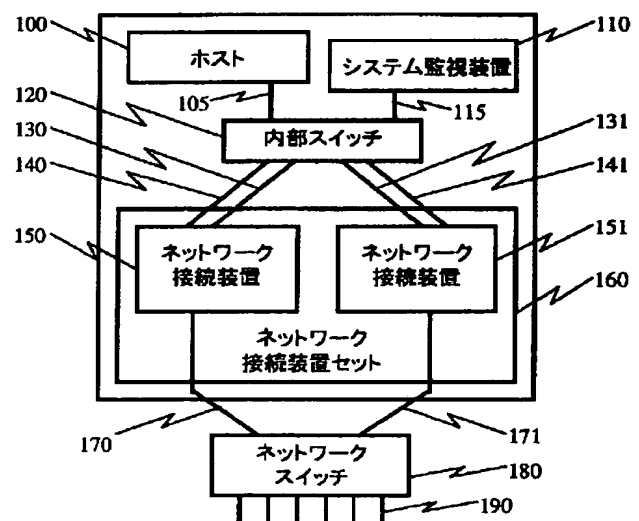
## (54) 【発明の名称】 多重化したネットワーク接続装置システム

## (57) 【要約】

【課題】 二重化したネットワーク接続装置を正常稼動時は負荷分散して性能向上に利用し、障害時にはホスト計算機のソフトの介入なしに可用性を維持する。

【解決手段】 受信処理ではネットワークスイッチ180に各ネットワーク接続装置150、151を接続し、全く同じ受信データを受信し、その内容からそれぞれでホストに上げるべき受信データを選択し、全体として受信データの脱落と重複がないようにし、障害発生時には残りのネットワーク接続装置が選択する受信データを動的に変更して、全体として受信機能を失わないようにする。

図1



## 【特許請求の範囲】

【請求項1】ホスト計算機をネットワークに接続するネットワーク接続手段であって、内部スイッチを備え、この内部スイッチではホスト計算機と1台のネットワーク接続装置として接続し、複数のネットワーク接続装置を備え、各ネットワーク接続装置は前記内部スイッチに接続し、各ネットワーク接続装置はそれぞれネットワークスイッチインタフェースを備え、このネットワークスイッチインタフェースによってネットワークスイッチに接続し、前記ホスト計算機からの送信要求の処理では、前記内部スイッチが送信要求を前記複数のネットワーク接続装置のいずれか1つに転送し、このネットワーク接続装置が転送された送信要求を実行し、前記ホスト計算機への通信データの処理では、前記ネットワークスイッチが通信データを前記複数のネットワークスイッチインタフェースを介して前記複数のネットワーク接続装置のそれぞれに送り、複数のネットワーク接続装置がこの通信データを受信すると、そのうちの1台のネットワーク接続装置のみがこの受信データをホスト計算機に転送することを特徴とする多重化したネットワーク接続装置システム。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】本発明は計算機システムをネットワークに接続する通信装置に係り、特にその良好な信頼性と性能を実現する装置構成と制御方法に関する。

## 【0002】

【従来の技術】LAN (Local Area Network) にワークステーション、パソコン等の計算機を接続する場合、計算機はNIC (Network Interface Card) 等のネットワーク接続装置を介して接続する。

【0003】通信ポイントを識別するために、物理レイヤの通信では各NICに付与された物理アドレスを用いる。イーサネット（登録商標）の場合、物理アドレスはMAC (Media Access Control) アドレスと呼ばれ、各NICに対してユニークな物理アドレスを製造者が付与して出荷している。

【0004】ネットワークレイヤの通信では、インターネットに代表されるTCP/IP通信プロトコルに基づくネットワークではIPアドレスを用いる。IPアドレスは1つのNICに対応して1つ付与し、計算機上のソフトウェアが用いる。

【0005】1台の計算機には通常1つのネットワーク接続装置を搭載し、従って1つのIPアドレスを持つ。但し、Douglas Comer著「Internetworking With TCP/IP Vol I: Principles, Protocols, And Architecture」の4.4節にある通り、1台の計算機が複数の、例えば2つのネットワーク

接続装置を搭載することも可能であり、この場合は計算機は各ネットワーク接続装置に対応して2つのIPアドレスを持つ。IPアドレスとネットワーク接続装置の関係は計算機上のソフトウェアによって管理する事ができる。より厳密には複数のネットワーク接続装置はそれぞれのIOポートに接続され、各ネットワーク接続装置をドライブするにはIOポートを区別して制御する必要があり、計算機上のソフトウェアはIPアドレスとIOポートの関係を管理する。

【0006】計算機が他の計算機と通信を行うには、通信先の計算機のIPアドレスと物理アドレスを知る必要がある。ソフトウェアは通信先のIPアドレスは外部から、例えばユーザの入力により与えられるが、このIPアドレスに対応する物理アドレスを知るためにはARP (Address Resolution Protocol) を用いる。ARPではIPアドレスを含むパケットをブロードキャストし、該当する計算機が対応する物理アドレスを返すことにより、物理アドレスを計算機上のソフトウェアに知らせる事ができる。

【0007】ネットワーク、およびネットワークを利用したアプリケーションが広く使われるようになってきている。計算機システムを互いにネットワークでつなぎ、ネットワークを経由して他の計算機が提供するサービスを利用したり、他の計算機上にある資源を利用したりするのが常態となっている。

【0008】他の計算機にネットワークを通じて様々なサービスを提供する計算機はサーバと呼ばれる。サーバとは広い意味での計算機、即ちネットワーク接続装置等やその上で稼動するソフトウェアも含んだ全体を指す。これに対し、サーバ内の構造を考える時は、演算装置や記憶装置を含む計算機本体はネットワーク接続装置等と対比してホストと呼ばれる。

【0009】上記のような状況下ではサーバのサービスの可用性が重要になってくる。サーバのサービスの可用性はサーバ自体の信頼性・可用性の他に、ネットワークの信頼性・可用性に大きく依存する。ネットワークの信頼性とは、ネットワーク上で通信されるデータの信頼性、即ち、パケット内のデータエラーの有無、パケット欠落の有無、およびそれらの回復メカニズムのことを言う。元来ネットワークは外部の影響を受けやすく、通信路上でのビットの欠落、通信装置の過負荷に伴うパケットの欠落といったエラーが一定の確率で発生する。

【0010】そのため各通信階層のプロトコルはこのようなエラーの発生を想定して設計されている。例えばTCP (Transmission Control Protocol) では、パケット内のデータエラーやパケットの欠落を検出する手段、欠落を検出した時は通信データを再送するなどの手段を含んでいる。またUDP (User Datagram Protocol) ではパケット内のデータエラーの検出を可能としているも

の、パケットの欠落についてはUDPを使用する上位のソフトウェアがそれぞれ対処するものとしてUDPとしては何もしない。

【0011】一方、ネットワークの可用性とは、通信手段の障害の有無とその回復メカニズムのことを言う。例えば通信を中継する装置の故障により通信路が切断されるような場合はネットワークの可用性の問題と考えられる。またサーバに装備されたネットワーク接続装置に障害が生じて、サーバそのものは正常であるにも関わらず、ネットワークに対するアクセスができなくなるような場合もネットワークの可用性の問題と考えられる。通信路の切断の問題は通信経路を複数設ける事により回避できる。

【0012】一方、ネットワーク接続装置の障害の問題はネットワーク接続装置を1つだけ搭載した構成では自動的な回避は不可能であり、例えば障害の発生したネットワーク接続装置を正常なネットワーク接続装置と交換する必要がある。

【0013】ネットワーク接続装置の障害に対する可用性を実現するために、ネットワーク接続装置の障害をサーバの障害と考えて、サーバごと交替させるというアプローチがある。また、サーバに2つのネットワーク接続装置を搭載し、一方に障害が発生した時に自動的に他方のネットワーク接続装置に切り替えて通信を回復することが考えられる。サーバが高価で信頼性が高いならば、ネットワーク接続部分の高可用性のためにこのような構成がありうる。

【0014】いずれの場合でも、前記のように、各ネットワーク接続装置にはそれぞれ固有のMACアドレスが付随し、またサーバのソフトウェアはそれぞれに固定したIPアドレスを割り当てて使用すると、サーバのソフトウェアやサーバの通信相手にネットワーク接続装置の障害を知られずにネットワーク接続装置を切り替える事はできないという問題がある。

【0015】特開平6-59924号公報「2重化システムの切替方式」では、運用系と待機系の2つのサーバがあり、運用系に障害が発生した時は各クライアントに伝文を送り、運用系切り替えのためのコネクション確立を行う。この方式ではサーバ・クライアントともに障害発生を認識し、対応した処理を行わなければならない。

【0016】特開平9-326810号公報「障害時のコネクション切り替え方法」では、運用系と待機系の2つのサーバがあり、それぞれのネットワーク接続装置(LANアダプタ)には同じIPアドレスを用い、運用系のネットワーク接続装置を有効に、待機系のネットワーク接続装置を無効に設定しておく。運用系に障害が発生した時はサーバの交代とともにそれぞれのネットワーク接続装置の有効・無効を切り替える。

【0017】クライアントは一旦コネクションの切断を検出するものの、サーバと新たなIPアドレスをやりと

りすることなく、それまでと同じIPアドレスに再接続要求を出すのみで新しい運用系サーバとコネクションを張ることができる。MACアドレスは再接続の際に解決されるので、2つのネットワーク接続装置に同じMACアドレスを設定しておく必要は必ずしもない。この方式では端末側の障害処理の負担が軽減されているものの、サーバ側のソフトウェアは依然として2つのネットワーク接続装置を区別した取り扱いを行う必要がある。

【0018】ネットワーク接続装置の障害時にサーバごと交替させる方式では、サーバ回復処理を行う必要があり、サーバのソフトウェアの介入は避けられない。1サーバに2つのネットワーク接続装置を搭載してネットワーク接続装置の交替を行う構成でも、サーバのソフトウェアが一方のネットワーク接続装置の障害を検出すると、他方のネットワーク接続装置に元のIPアドレスを割り当てて利用することができるが、この時、ソフトウェアはIPアドレスとIOポートの対応関係を変更しなければならない。

【0019】また、これらの技術では、正常運用中は待機系が丸々遊んでしまい、コストパフォーマンス上の不満がある。

【0020】

【発明が解決しようとする課題】上記の問題点に鑑み、本発明の目的は、ネットワーク接続装置の可用性向上に関し、サーバのネットワーク接続装置に障害が発生してもサーバのソフトウェアにその事実を極力意識させず、複雑な回復処理のような介入なしに交替・回復が可能で、かつ多重に備えたネットワーク接続装置をネットワーク性能の高性能化のためにも利用することが可能なコストパフォーマンスに優れた多重化したネットワーク接続装置と、その制御方法を提供するという課題を解決しようとするものである。

【0021】

【課題を解決するための手段】上記課題を解決するための本発明による一つの構成では、ホストの1つの入出力ポートに内部スイッチを設け、これに複数のネットワーク接続装置を搭載し、これらのネットワーク接続装置を1つのネットワークスイッチにそれぞれ接続する。ホストからの送信の時は内部スイッチにおいていずれのネットワーク接続装置を使用するかを決定する。複数のネットワーク接続装置には同じMACアドレスを設定する。またサーバが受信する時はネットワークスイッチは同じ内容を各ネットワーク接続装置に同時に送るように構成し、各ネットワーク接続装置において受信したデータをさらにホストに伝達するか否かを決定する。

【0022】1つのネットワーク接続装置に障害が発生した場合は、内部スイッチが障害の発生を認識し、障害の発生したネットワーク接続装置には送信データを送り出さないようにする。また受信データの場合、障害の発生したネットワーク接続装置は受信データをホストに伝

達せず、他のネットワーク接続装置が受信データをホストに伝達するようにする。これらの構成を可能ならしめるために、ネットワーク接続装置には障害発生を検出した時には内部スイッチに障害発生を報告すると同時に、障害状態に移行する障害処理機構を設ける。

【0023】以上により、ネットワーク接続装置の障害に対する可用性が実現され、内部スイッチで障害ネットワーク接続装置を認識するため、サーバのソフトウェアはネットワーク接続装置が二重であること、ネットワーク接続装置の障害発生時にも特別な処理を行う必要はない。また平常時は複数のネットワーク接続装置がそれぞれ利用されるため、性能的にも優れたものになる。

【0024】

【発明の実施の形態】以下、図を用いて本発明の実施の形態を説明する。ここでは簡単のために主に2台のネットワーク接続装置によって二重化されている場合について説明するが、本発明はこれに限るものでなく、3台以上の構成で多重化されている場合でも有効である。また多重化されたネットワーク接続装置によってなる構造をここではネットワーク接続装置セットと呼ぶことにする。ネットワーク接続装置における通信データの流れは、ホストによる出力、すなわち送信と、ホストによる入力、すなわち受信の2方向がある。

【0025】図1は本発明の概要を表わすブロック図である。ホスト100は独立した計算機であり、オペレーティングシステム、アプリケーションソフトウェア等が稼動する。ホスト100と本発明にかかるネットワーク接続装置セット160は1つの内部スイッチ120を介して接続されている。ホスト100と内部スイッチ120はホスト入出力インタフェース105によって接続されている。内部スイッチ120とネットワーク接続装置セット160内の各ネットワーク接続装置150、151はそれぞれ通信データと制御信号を通すネットワーク接続装置インタフェース130、131で接続され、更に両者の間にはネットワーク接続装置状態管理インタフェース140、141が設けられている。

【0026】ネットワーク接続装置状態管理インタフェース140、141は必ずしもネットワーク接続装置インタフェース130、131と独立した接続ではなく、ネットワーク接続装置インタフェース130、131を利用した通信によって実現することも可能である。内部スイッチ120はまたログインタフェース115により、システム監視装置110と接続されている。システム監視装置はネットワーク接続装置に障害が発生した場合に障害のあったネットワーク接続装置の状態情報の保管、障害を起こしたネットワーク接続装置の交換作業の支援などを行う。

【0027】ネットワーク接続装置150、151はそれぞれネットワークスイッチインタフェース170、171により、本発明にかかるネットワークスイッチ18

0に接続されている。ネットワークスイッチ180は通信インタフェース190により通信ネットワークに接続されている。ネットワーク接続装置セット160内の各ネットワーク接続装置150、151は同じMACアドレスを持つ。MACアドレスはネットワークインタフェースについて製造者が一意的な番号をつけることになっているが、本実施例ではネットワーク接続装置セット160で1つのネットワークインタフェースと考えられ、このような構成は正当である。

【0028】図2は本発明にかかる内部スイッチ120の構造を示したブロック図である。ホスト100からの送信データはホスト入出力インタフェース105を経由してセクタ200に入る。セクタ200ではこの出力データを複数あるネットワーク接続装置インタフェース130、131から一つを選択して転送する。ネットワーク接続装置インタフェースの選択はネットワーク接続装置状態管理装置220からの選択制御インタフェース230を用いて行う。

【0029】ネットワーク接続装置状態管理装置220は各ネットワーク接続装置の状態情報をネットワーク接続装置状態管理インタフェース140、141から得ることができ、これを用いて選択制御インタフェース230に出力する。各ネットワーク接続装置の状態情報としてはその稼動状況、障害状況などがある。また障害発生時に保守情報を保管するためにログインタフェース115を持ち、システム監視装置に接続されている。

【0030】尚、図2において、受信データは複数あるネットワーク接続装置インタフェース130、131のいずれかから入り、セクタ200を経由してホストの入力となる。ネットワーク接続装置インタフェース130、131からの入力がセクタ200において衝突することがあるので、セクタ200ではその調停制御を行う。

【0031】図3は本発明にかかるネットワーク接続装置150の構造を示したブロック図である。ホスト100から内部スイッチ120を経て出力される送信データはネットワーク接続装置インタフェース130を経てネットワーク接続装置150のホストインタフェース300に入る。ホストインタフェース300は内部スイッチ120のセクタ200とのデータ転送制御を行い、送信データを内部接続310を介して転送制御装置320に送り込む。

【0032】転送制御装置320は内部接続330を介してこのデータをネットワークインタフェース340に送る。転送制御装置320、およびネットワークインタフェース340ではイーサネットのフレームを構成してネットワークスイッチインタフェース170に送出する。このネットワーク接続装置のMACアドレスはMACアドレス格納装置360に格納されており、転送制御装置320により設定・変更ができる。ネットワークイ

インタフェース 340 で送出するフレームには MAC アドレス格納手段 360 に設定された MAC アドレスが付加される。

【0033】一方、ネットワークスイッチインタフェース 170 から到来する受信データはフレームに付加された MAC アドレスとこのネットワーク接続装置の MAC アドレス 360 との一致を条件にネットワークインタフェース 340 に取り込まれ、転送制御装置 320 に送られる。転送制御装置 320 では受信データを更にホストインタフェース 300 を介してホストまで転送するか否かを定める。

【0034】この制御は受信したデータに依存し、さらにネットワーク接続装置状態管理装置 220 からの選択制御情報 350 を用いて行う。選択制御情報 350 はネットワーク接続装置状態管理インタフェース 140 からもたらされる。転送制御装置 320 では受信データのホストへの転送を行わないと決めた場合、この受信データを棄却する。転送が棄却かの決定をフレームの受信後、速やかに行うことによって、ネットワーク接続装置の負荷を抑えることができる。

【0035】図 4 はネットワークスイッチ 180 の構造を示すブロック図である。ネットワークスイッチインタフェース 420～427 はそれぞれ、図には示されていない計算機、ルータ、ネットワークスイッチなどとルーティングスイッチ 410 を接続している。ここでは各ネットワークスイッチインタフェース 420～427 は双方向の通信線である。例えばルーティングスイッチ 410 は、ネットワークスイッチインタフェース 420 からの入力をその宛先アドレスに応じて例えばネットワークスイッチインタフェース 427 に出力する。

【0036】分配器 400 はネットワークスイッチ 180 の内部にあるネットワークスイッチインタフェース 423 に接続されており、さらに 2 本のネットワークスイッチインタフェース 170、171 によって本発明にかかるネットワーク接続装置 150 のネットワークインタフェース 340、およびネットワーク接続装置 151 のネットワークインタフェースにそれぞれ接続されている。分配器 400 は次のように構成する。フレームが内部インタフェース 423 に出力されると分配器 400 はネットワークスイッチインタフェース 170、171 の双方に同じ内容のフレームを同時に出力する。

【0037】またネットワークスイッチインタフェース 170、171 からの通信パケットはそれぞれ独立に分配器 400 により内部インタフェース 423 に送出され、ルーティングスイッチ 410 により転送される。ここでは分配器 400 はネットワークスイッチ 180 の内部に描かれているが、外部に引き出してから取り付けても良く、また本発明にかかるネットワーク接続装置セット 160 の内部に設けても良い。

【0038】図 5 はネットワーク接続装置状態管理装置

220 の制御に用いるデータ構造を示している。ネットワーク接続装置状態管理テーブル 500 はネットワーク接続装置に対応する行を持ち、それぞれネットワーク接続装置識別フィールド 510、ネットワーク接続装置状態フィールド 520、受信データ選択条件フィールド 530、およびネットワーク接続装置負荷指標フィールド 540 を持つ。これらの情報はネットワーク接続装置状態管理インタフェース 140、141、ないし 142 により各ネットワーク接続装置から得られる情報を用いて作成・更新され、選択制御インタフェース 230 の出力に反映される。尚、ここでは説明のためネットワーク接続装置状態管理インタフェースは 3 本描かれている。これは描かれていないネットワーク接続装置が 3 台存在することを示している。

【0039】ネットワーク接続装置識別フィールド 510 は接続されている、あるいは接続可能なネットワーク接続装置を識別し、指定する情報を格納する。図では例として「0」「1」「2」「3」が格納されている。ネットワーク接続装置状態フィールド 520 は対応するネットワーク接続装置が正常状態か故障状態か、あるいは実装されているかなどを表わす。

【0040】図では「正常」「故障」「不在」が表示されている。「正常」は該当するネットワーク接続装置が正常に動作していることを示す。「故障」は該当するネットワーク接続装置に故障が発生しており、使用できないことを表わしている。「不在」は該当するネットワーク接続装置が搭載されておらず、使用できないことを表わしている。

【0041】これらの情報を用いて、故障状態のネットワーク接続装置、あるいは実装されていないネットワーク接続装置には送信データを振り向けないよう、選択制御インタフェース 230 の出力を作成し、セレクト 200 を制御する。受信データ選択条件フィールド 530 については後から説明する。

【0042】図 6 は 1 つのネットワーク接続装置の負荷の程度を表わす負荷指標をネットワーク接続装置状態管理装置 220 に報告する際に送られる負荷指標情報の構造を示している。負荷指標情報 600 はネットワーク接続装置状態管理テーブル 500 の各行と類似の構造を持ち、ネットワーク接続装置識別フィールド 610、ネットワーク接続装置状態フィールド 620、およびネットワーク接続装置負荷指標フィールド 640 を持つ。

【0043】ネットワーク接続装置識別フィールド 610 は該当するネットワーク接続装置を示す。ネットワーク接続装置状態フィールド 620 の値は「正常」とし、このネットワーク接続装置が正常に動作していることを示すと同時にこの情報が負荷指標情報であることを表示する。

【0044】ネットワーク接続装置負荷指標フィールド 640 には各ネットワーク接続装置の負荷状態を表わす

データが格納されている。ここでは負荷状態表示方法の一例として負荷状態を0, 1, 2, 3の4段階に分けて表示している。数字は大きいほど負荷が高いものとし、例えば過去1秒間のネットワーク接続装置のビジー率が0~15%ならば0、15~30%ならば1、30~50%ならば2、50~100%ならば3をそれぞれ負荷指標として与えるものとする。

【0045】負荷指標情報600のネットワーク接続装置状態管理装置220への報告のタイミングとしては、1つの送信処理ないし受信処理が終了した時点で報告するものが考えられる。しかし、この報告は必ずしも送信処理・受信処理と同時に行為される必要はなく、送信・受信をいくつか処理する毎に報告するもの、あるいは一定時間、例えば1秒おきに報告することなどが考えられる。一定時間に報告させる方式とするとこれを対応するネットワーク接続装置の時間監視制御と兼用して利用することができる。

【0046】負荷指標情報600を受け取るとネットワーク接続装置状態管理装置220ではそれを元にネットワーク接続装置状態管理テーブル500の該当する行を更新する。該当する行は負荷指標情報600のネットワーク接続装置識別フィールド610とネットワーク接続装置状態管理テーブル500のネットワーク接続装置識別フィールド510の一致を調べることにより定まる。このようにして定まった行のネットワーク接続装置負荷指標フィールド540にネットワーク接続装置負荷指標フィールド640の値をコピーする。

【0047】このようにして作成したネットワーク接続装置状態管理テーブル500の情報を用いて、セレクト200がどのように送信データを送出するネットワーク接続装置を選択するかについて述べる。ネットワーク接続装置状態フィールド520の値が「正常」であるものから選択し、「故障」あるいは「不在」であるものは使用しない。セレクトが使用を試みるネットワーク接続装置が「正常」であっても、受信処理中であつたり、別の送信処理中である場合はそのネットワーク接続装置を直ちに使用することはできない。この時は残りの正常なネットワーク接続装置から使用可能なものを使用する。

【0048】このような場合、どのネットワーク接続装置から使用を試みるかを決めるために、セレクト200は選択制御インタフェース230を介してネットワーク接続装置負荷指標フィールド540を参照し、その値の小さいものから順に使用を試みるものとする。但し、その値が等しい時はあらかじめ決めておいた順序に従って使用を試みるものとする。このように構成することにより、受信処理で負荷の高くなっているネットワーク接続装置には送信処理をやらせないようになり、負荷の分散を図ることができる。

【0049】図12と図13により上に述べた制御を実現する一実施例を説明する。図12のネットワーク接続

装置状態管理テーブル1200は図5のネットワーク接続装置状態管理テーブル500とは異なった状態を示している。ここではネットワーク接続装置識別子ID=0, 1, 3の3つのネットワーク接続装置が正常に稼働しており、対応する負荷指標1241, 1242, 1244がそれぞれ3, 3, 1となっている。

【0050】このときネットワーク接続装置状態管理装置220は負荷指標の小さい順に従って送信装置選択情報1290を作成する。送信装置選択情報1290は選択すべきネットワーク接続装置の識別子の並びがそれぞれ1291, 1292, 1293に格納する。ID3のエントリ1294には本来ならば正常なネットワーク接続装置の識別子が格納されるべき所であるが、ID=2のネットワーク接続装置が故障中であるため、候補のネットワーク接続装置が無いことが「X」によって表現されている。

【0051】図13は送信装置選択情報1290を用いたセレクト200におけるネットワーク接続装置の選択処理を示したフローチャートである。ステップ1305でまず送信装置選択情報1290をネットワーク接続装置状態管理装置220から得る。これは、予めネットワーク接続装置状態管理装置220から与えておいた送信装置選択情報1290をセレクト内の記憶装置に格納しておく、あるいはステップ1305の中でネットワーク接続装置状態管理装置220に問い合わせる、といった方法がある。前者の方が送信要求が出てからの処理が短く、より良い性能が選られると考えられる。

【0052】ステップ1310では送信装置選択情報1290をサーチするのに用いるエントリポインタ1390を先頭のエントリ1291に位置づける。ステップ1320ではエントリポインタ1390の指すエントリに有効なネットワーク接続装置識別子があるか否かを調べる。先頭のエントリには有効な識別子が入っており、YESの分岐に進む。ステップ1340ではこのエントリから選択候補のネットワーク接続装置の識別子を取り出す。

【0053】次にステップ1350ではこのネットワーク接続装置が現在使用可能か否かを調べる。このネットワーク接続装置が使用可能ならばYESに分岐し、ステップ1360でこのネットワーク接続装置を送信に用い、選択処理1300は完了する。使用不可ならばNOに分岐に進む。使用できない状況としては他の送信処理、あるいは受信処理中である場合がある。

【0054】使用できない時はステップ1370でエントリポインタ1390を一つ進め、帰ループ1380に従って、ステップ1320に戻る。改めてエントリポインタ1390の指すエントリを調べると有効な識別子があればこれについて同様に使用を試みる。例えば図13のエントリポインタ1390がエントリ1294を指しているとする、そのエントリは無効であるからステ

ップ1320からNOに分岐し、ステップ1330に進む。ステップ1330では全てのネットワーク接続装置の候補がビジーであるので、装置ビジー処理を行う。装置ビジー処理としては、送信要求元に装置ビジーを応答する、あるいは一定期間待って、もう一度、ネットワーク接続装置の選択を試みる、などが考えられる。

【0055】次にネットワーク接続装置が受信した受信データの選択、即ちその受信データをホストに転送するか否かを定める方式について説明する。本実施例では受信データのもつデータからホスト転送するネットワーク接続装置を1つ決める。別の言い方をすると各ネットワーク接続装置で受信データから適当なデータ（選択データと呼ぶことにする）を取り出し、その値に応じてそのネットワーク接続装置がその受信データをホストに転送するか否かを決定する。正常な複数のネットワーク接続装置のいずれか1つのみが必ずホストに転送することを決定する。また、複数のネットワーク接続装置からホスト転送を行うものがあるべく一様に選ぶようにして、負荷分散を図るようにする。

【0056】受信データから選択データを取り出す方法を説明する。図8、図9、図10、図11は受信データのデータ構造を示している。各フィールドの詳細はDouglas Comer著の前掲書を参照のこと。図8はイーサネット上のUDP/IP、あるいはTCP/IPで通信されるフレーム800の構造を示している。フレームヘッダ、IPヘッダに続いて、UDPヘッダあるいはTCPヘッダがある。図9はIPパケットヘッダ900の構造を示している。

【0057】図10はUDPヘッダ1000の構造を示しており、UDPパケットのチェックサムフィールド1040がある。図11はTCPヘッダ1100の構造を示している。UDPヘッダのUDPチェックサム1040と、TCPヘッダの通し番号フィールドの下位16ビット1140がフレーム800上で同じ位置にある。このフィールドのデータとIPヘッダ900の中の送信元IPアドレスの下位16ビットとの排他的論理和を選択データとする。より具体的にはその下位3ビットの値を用いるものとする。

【0058】UDPチェックサムフィールドにはパケット内容に依存しており、内容が異なれば選択データの値は散らばるであろう。またWEBサーバへのアクセスのように、内容が同じものが集中するような場合や、UDPのチェックサムを使用していない場合（UDPではこれはユーザのオプションである）はチェックサムの値が偏る可能性が高いが、送信元毎に異なるIPアドレスを計算に組み入れているため、得られる値の分布は散らばるものと期待できる。またTCPの通し番号は毎回異なる値となる。いずれの場合も下位3ビットの値の分布は一様になり、ネットワーク接続装置の間の受信データ処理に関する負荷は分散されるものと期待できる。

【0059】上記ようにして受信データから得られた選択データの値0～7を用いて、この受信データをホストに転送するか否かを定める方法を説明する。ネットワーク接続装置150には選択制御情報350がある。今その値が{0, 2, 4, 6}（集合値）とする。選択データがこの集合の中のいずれかに等しい時、このネットワーク接続装置はこの受信データをホストに転送することに決定する。

【0060】ネットワーク接続装置における受信データの選択が上記のように行われる時、複数のネットワーク接続装置による受信処理を次のように制御する。ネットワーク接続装置状態管理テーブル500の各行の受信データ選択条件フィールド530には該当するネットワーク接続装置の選択制御情報350に設定すべき値を格納する。

【0061】正常なネットワーク接続装置が複数台ある時、それらには選択データのとりうる値の集合{1, 2, ..., 7}を分割したものを割り当てる。即ち、任意の2つのネットワーク接続装置については必ず異なる選択データが割り当てられ、かつすべての選択データの値はいずれかのネットワーク接続装置に割り当てられている。例えば、図5のネットワーク接続装置状態管理テーブル500には識別子が0, 1のネットワーク接続装置についてこの値がそれぞれ{0, 2, 4, 6}、{1, 3, 5, 7}である状態が示されている。

【0062】また図12のネットワーク接続装置状態管理テーブル500には同様に{0, 2, 4}、{1, 5, 6}、{3, 7}が3つのネットワーク接続装置に割り当てられている様子が示されている。このように割り当てることにより、「正常な複数のネットワーク接続装置のいずれか1つのみが必ずホストに転送する」という前記の条件を満たされる。

【0063】ネットワーク接続装置状態管理装置220は正常なネットワーク接続装置がどれであるかが与えられた時、ネットワーク接続装置状態管理テーブル500の受信データ選択条件フィールド530を自動的に生成する。例えば、正常なネットワーク接続装置を並べ、選択データの取り得る値、本実施例では0, 1, ..., 7を順次割り当てれば良い。さらに受信データ選択条件フィールド530に設定された値を各ネットワーク接続装置の選択制御情報350に設定する。ネットワーク接続装置状態管理インタフェース140, 141, ないし142を通じて各ネットワーク接続装置に対してこの設定はなされる。

【0064】次にネットワーク接続装置に障害が発生した場合の制御方式について説明する。図7は1つのネットワーク接続装置で障害を検出した時にネットワーク接続装置状態管理装置220に送られる障害検出情報700の構造を示している。障害検出情報700はネットワーク接続装置識別フィールド710、ネットワーク接続

装置状態フィールド720を含む。ネットワーク接続装置識別フィールド710は該当するネットワーク接続装置を示す。ネットワーク接続装置状態フィールド720の値は「故障」とし、このネットワーク接続装置に障害が発生したことを示すと同時に、この情報が障害検出情報であることを表示する。

【0065】図14はネットワーク接続装置状態管理装置220が障害検出情報700を受けた時の処理1400を示すフローチャートである。ネットワーク接続装置状態管理装置220は、ステップ1410において、障害検出情報700のネットワーク接続装置識別フィールド710から障害を起こしたネットワーク接続装置の識別子を得る。ステップ1420では、得られた識別子に該当するネットワーク接続装置状態管理テーブル500のエントリのネットワーク接続装置状態フィールド520の値を「故障」に変更する。

【0066】ステップ1430ではこの変更を選択制御インタフェース230によりセクタ200にも反映させ、セクタ200における障害の起きたネットワーク接続装置を送信データの転送のために選択することのないようにする。ステップ1440では、該当するエントリの受信データ選択条件フィールド530をクリアしてこのネットワーク接続装置はいかなる受信データもホストに転送しないことを表示する。

【0067】ステップ1450では他のネットワーク接続装置の受信データ選択条件フィールド530を書き換えて、残りのネットワーク接続装置の間ですべての受信データが重複なくいずれかのネットワーク接続装置によってホストに転送されるように設定し直す。この処理はネットワーク接続装置状態管理装置220により自動的に行われることはすでに述べた通りである。

【0068】これにより、障害の発生したネットワーク接続装置がホストに転送するはずだった受信データは他のネットワーク接続装置によってホストへ転送されるようになる。更にステップ1460では保守情報として障害を起こしたネットワーク接続装置の状態情報をログインタフェース115に出力する。この情報はシステム監視装置110によって保存される。続いてステップ1470では障害を起こしたネットワーク接続装置の交換保守を可能とするためにこのネットワーク接続装置の電源を切断する。ステップ1480で一連の障害処理の終了をランプ点灯などにより表示する。

【0069】図15は上記の処理で交換可能となった障害の発生したネットワーク接続装置を実際に交換し、再び使用を開始する為の方法を説明したフローチャートである。複数の処理の流れが相互に関連している。

【0070】第1の処理の流れはシステム監視装置110によるネットワーク接続装置への電源投入処理1500である。システム監視装置はステップ1510で電源投入を実行する。矢印1590はこの電源投入を契機に

第2の処理であるネットワーク接続装置の稼働時回復処理1520が起動されることを表わしている。

【0071】ネットワーク接続装置は内部的な初期化処理を実行し、ステップ1530にてネットワーク接続装置状態管理装置220に等ネットワーク接続装置が新たに起動されたことを報告する。矢印1592はこの報告を契機に第3の処理であるネットワーク接続装置の復帰処理1540が起動されることを表わしている。ネットワーク接続装置はステップ1532にてこの後のネットワーク接続装置からの要求を待つ。

【0072】ステップ1550で、ネットワーク接続装置状態管理装置220は復帰処理対象のネットワーク接続装置に対して、このネットワーク接続装置セットに対して設定されているMACアドレスを設定する。矢印1594はこの設定要求の発行を表わしている。これによってネットワーク接続装置における要求待ち状態は解除され、ステップ1534に進み、指定されたMACアドレスをMACアドレス格納装置360に格納する。これによって、このネットワーク接続装置による送信処理は可能になる。ついでステップ1536にて再び要求待ち状態に入る。

【0073】ステップ1552で、ネットワーク接続装置状態管理装置220はネットワーク接続装置状態管理テーブル500の該当するエントリを作成する。例えば起動前のネットワーク接続装置状態管理テーブル500の状態が図12の1200の通りであったとする。すなわち、識別子が「2」のネットワーク接続装置が故障している。この「2」のネットワーク接続装置を回復、起動することにより、テーブルの状態は図12の1201のように変更する。すなわち、状態は「正常」とし、選択条件を4台の正常なネットワーク接続装置にそれぞれ図のように割り付け直し、さらに負荷指標は「1」、すなわち一番負荷の軽い状態とする。

【0074】ステップ1554で、ネットワーク接続装置状態管理装置220は、セクタ200において使用可能なネットワーク接続装置も送信処理に使用されるよう、選択制御インタフェース230を通じて使用可能ネットワーク接続装置とその優先順序を設定する。これにより、送信データが回復中のネットワーク接続装置にも転送されるようになる。このネットワーク接続装置はすでに送信処理に関しては実行可能となっているため、実際に送信データを転送しても正しく送信処理を行うことができる。

【0075】セクタ200における選択メカニズムに応じて、上記の使用可能ネットワーク接続装置とその優先順序の実際の設定方法は変更を受ける。例えばセクタ200が毎回の送信処理で必ず、ネットワーク接続装置状態管理装置に使用可能なネットワーク接続装置を問い合わせるのならば、ネットワーク接続装置状態管理テーブルへの状態変更の反映だけで良い。しかし、セレク

タがネットワーク接続装置の状態を一定期間、キャッシュとして保持するならば、割込みをかけるなどして積極的にそれを更新させるべきである。

【0076】ステップ1556で、ネットワーク接続装置状態管理装置220は各ネットワーク接続装置に処理すべき受信処理の選択情報を通知する。この中には、以前から正常に稼動していたネットワーク接続装置と、今回、障害から回復したネットワーク接続装置も含まれている。矢印1596は障害から回復中のネットワーク接続装置への通知を表わしており、これによってネットワーク接続装置の要求待ち状態が解除される。この結果、ネットワーク接続装置はステップ1538にて受信処理も実行を開始する。以上でネットワーク接続装置の回復処理が完了する。

【0077】

【発明の効果】本発明によれば、二重化したネットワーク接続装置を正常稼動時は負荷分散して性能上有効に利用し、かつ障害発生時にはホスト計算機のオペレーティングシステム介入なしに可用性を維持することを可能となる。

【0078】内部スイッチでは各ネットワーク接続装置の負荷情報を元に送信処理を実行するネットワーク接続装置を選択するため、ネットワーク接続装置間の負荷バランスを図ることができ、それぞれのネットワーク接続装置をなるべく低負荷な状態で使用するため、安定、かつ良好な送信性能を実現することができる。

【0079】各ネットワーク接続装置では受信処理を分担して実行するため、受信処理における負荷バランスを図ることができる。受信処理の分担は到来した受信データの内容、特に上記実施例ではTCP/IPプロトコルのUDPプロトコルのUDPチェックサムのフィールドを利用しており、データ内容が異なればその値も変化する、自然と負荷分散を図ることができる。1つのクライアントからサーバにデータをバックアップするような場合、クライアントのIPアドレスのみを用いた場合では負荷分散されないが、データ内容を元にネットワーク接続装置を選択することにより、前記のような場合にも負荷分散を図ることができる。

【0080】実際には長大なUDPパケットは分割され、UDPチェックサムは先頭のフラグメントにしか付加されず、後続のフラグメントはUDPパケットのデータの内容を見ることになり、必ずしも理想的な負荷分散とならない可能性がある。しかし、前記のバックアップのような場合にはデータフィールドの内容により、十分な負荷分散が実現されるものと期待できる。

【0081】ネットワーク接続装置の障害発生時には障害のあるネットワーク接続装置をはずした構成で受信処理の分担等を自動的、かつシステム稼働中に動的に再設定すれば通信処理をそのまま続行でき、さらに障害を起こしたネットワーク接続装置を取り外し、正常なネットワーク接続装置と取り替えて通信処理に復帰させることができるため、システム全体での高い可用性が実現できる。

【0082】本発明ではネットワーク接続装置を二重化しているが、ネットワークスイッチ、及び内部スイッチの二重化には触れていない。しかし、これらのコンポーネントについては別的手段での二重化が可能であり、本発明と組み合わせれば、より信頼度の高い可用性が実現できる。

【図面の簡単な説明】

【図1】本発明の概要を表わすブロック図。

【図2】内部スイッチの構造を示すブロック図。

【図3】ネットワーク接続装置の構造を示すブロック図。

【図4】ネットワークスイッチの構造を示すブロック図。

【図5】ネットワーク接続装置状態管理装置の制御に用いるデータ構造を示す図。

【図6】ネットワーク接続装置を報告する負荷指標情報のデータ構造を示す図。

【図7】ネットワーク接続装置の障害検出情報のデータ構造を示す図。

【図8】受信データのデータ構造を示す図。

【図9】受信データのデータ構造を示す図。

【図10】受信データのデータ構造を示す図。

【図11】受信データのデータ構造を示す図。

【図12】ネットワーク接続装置状態管理テーブルの状態の例を示す図。

【図13】セレクトにおけるネットワーク接続装置の選択処理のフローチャート。

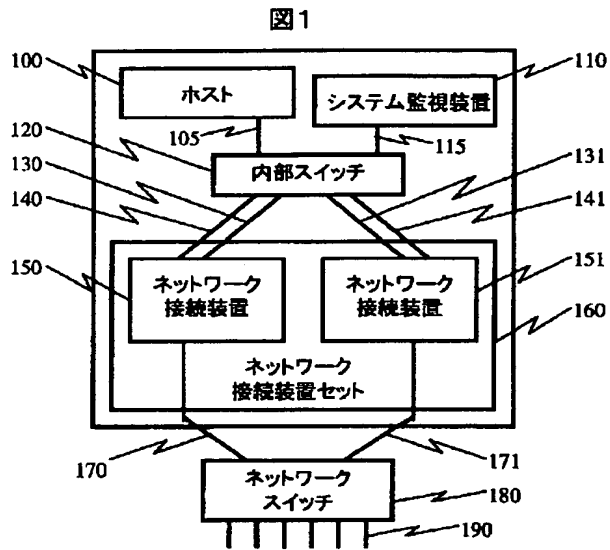
【図14】ネットワーク接続装置状態管理装置における障害検出情報処理のフローチャート。

【図15】ネットワーク接続装置の使用再開処理のフローチャート。

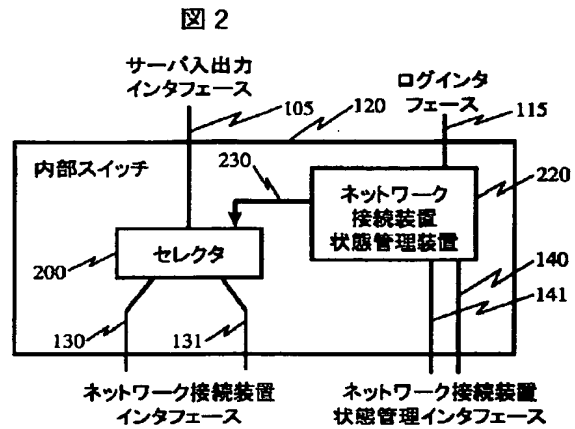
【符号の説明】

120…内部スイッチ、150、151…ネットワーク接続装置、180…ネットワークスイッチ、200…セレクト、220…ネットワーク接続装置状態管理装置、320…転送制御装置。

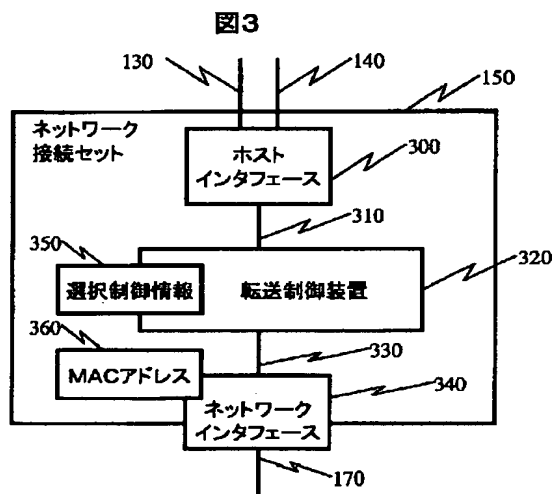
【図1】



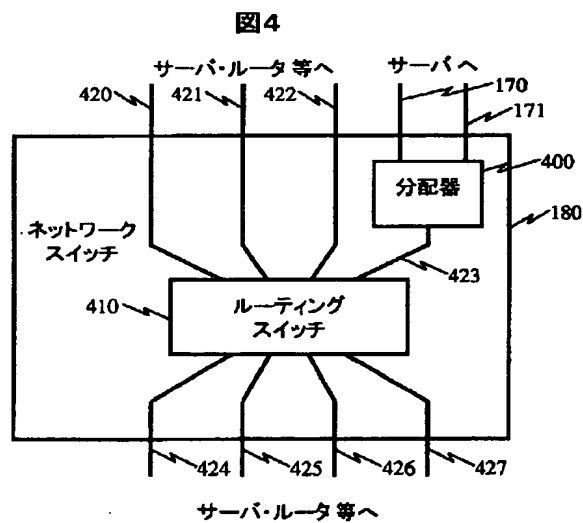
【図2】



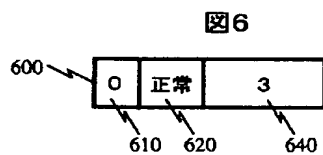
【図3】



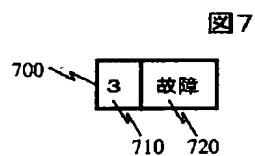
【図4】



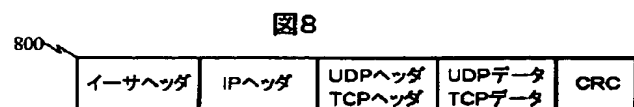
【図6】



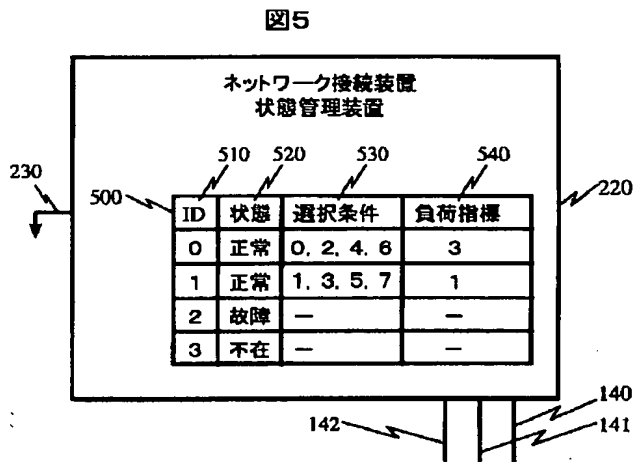
【図7】



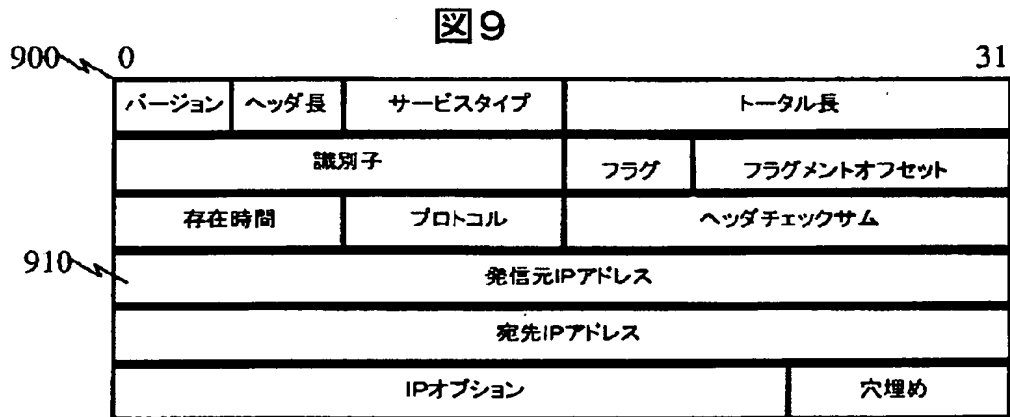
【図8】



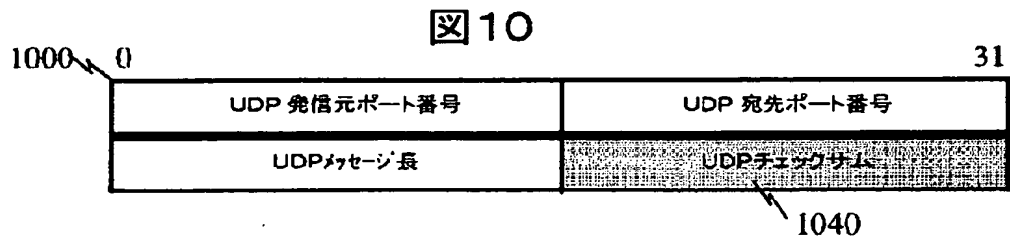
【図5】



【図9】

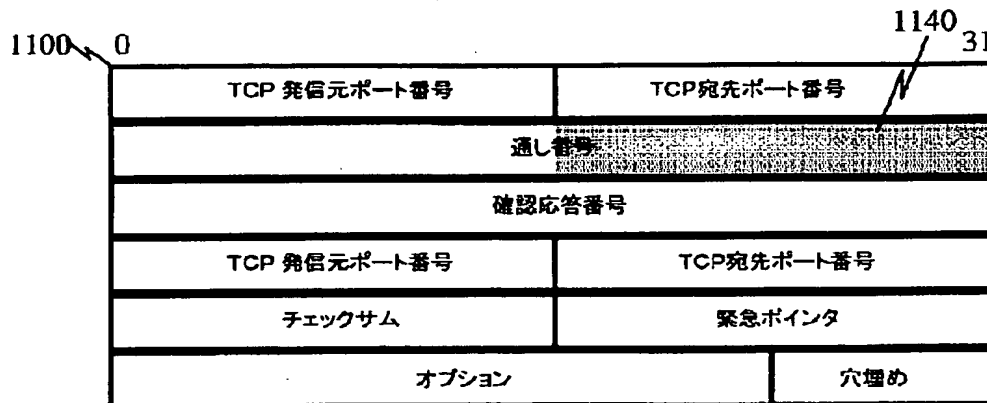


【図10】



【図11】

図11



【図12】

図12

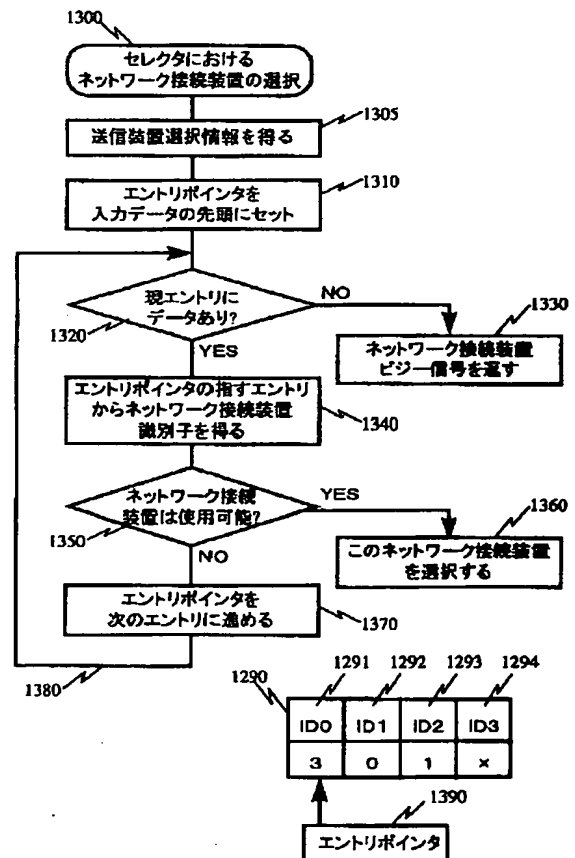
| ID | 状態 | 選択条件    | 負荷指標 |
|----|----|---------|------|
| 0  | 正常 | 0, 2, 4 | 3    |
| 1  | 正常 | 1, 5, 6 | 3    |
| 2  | 故障 | —       | —    |
| 3  | 正常 | 3, 7    | 1    |

| ID0 | ID1 | ID2 | ID3 |
|-----|-----|-----|-----|
| 3   | 0   | 1   | x   |

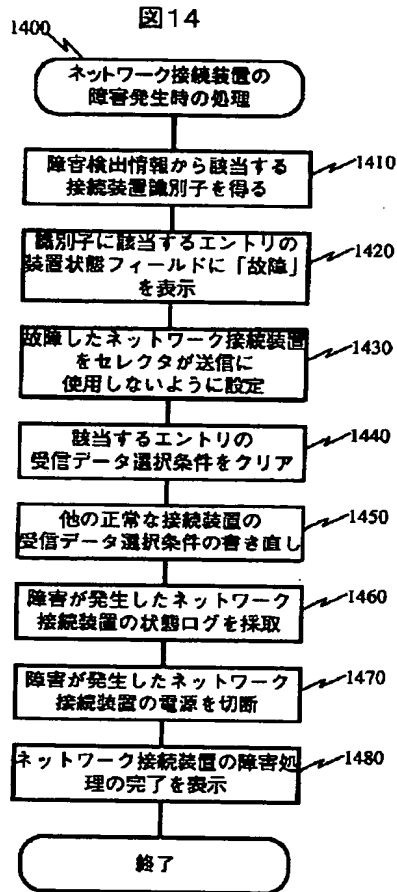
| ID | 状態 | 選択条件    | 負荷指標 |
|----|----|---------|------|
| 0  | 正常 | 0, 2, 4 | 3    |
| 1  | 正常 | 1, 5, 6 | 3    |
| 2  | 故障 | —       | —    |
| 3  | 正常 | 3, 7    | 1    |

【図13】

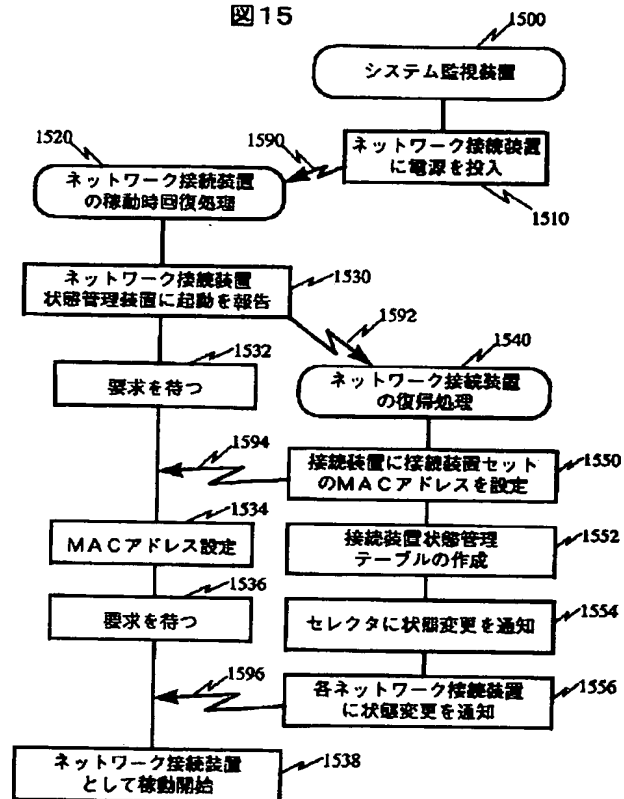
図13



【図14】



【図15】



フロントページの続き

(72) 発明者 富田 和良  
神奈川県秦野市堀山下1番地 株式会社日  
立製作所汎用コンピュータ事業部内

Fターム(参考) 5B089 GA31 GA32 KA12 MD02 ME04  
5K033 AA03 AA06 BA04 DA01 DB03  
DB16 DB17 DB20 EA03 EB06  
EB08  
5K035 AA02 AA03 BB03 CC05 DD01  
EE03 EE22 LL14